Случайные процессы. Прикладной поток.

Практическое задание 9*.

Соревнование на kaggle по предсказанию продаж
группы товаров одного из гипермаркетов Москвы.

Вам предоставляются данные по продажам одного из гипермаркетов Москвы с 1 августа 2015 года до 31 мая 2016 года. Фактически, данные представляют все чеки за данный период. Задача — как можно точнее предсказать продажи 20 различных товаров в течении следующего месяца, то есть с 1 по 30 июня 2016 года. Такая информация может помочь магазину правильно распределить, сколько товара в какое время нужно закупить. Ниже содержится информация о правилах. В целом правила максимально похожи на стандартные правила проведения соревнований на kaggle в рамках учебных курсов.

Ссылка: https://www.kaggle.com/c/dihtrandomprocess2017
Инвайт: https://www.kaggle.com/t/54b9040fba1e4d86bd325a883feb74b9

Дедлайн на kaggle: **29.12.2017 23:59**
Дедлайн по отправке решения на почту: **30.12.2017 23:59**
Дедлайны могут быть изменены в меньшую сторону!!!

Данные:
`train.csv` — файл, в котором записана информация о чеках. Содержит колонки:

- `TRANSACTIONID` — номер чека, может повторяться (это соответствует товарам, присутствующим в одном чеке);
- `ITEMID` — номер товара;
- `PRICE` — цена товара на момент продажи;
- `TRANSDATE` — дата покупки.

`items.csv` — файл с 20 товарами, продажи которых нужно предсказать.

Что является ответом?
Файл прогнозов формата `Id:Count`, в котором строка с номером $i * 30 + j$ соответствует прогнозу продаж товара i (нумерация с нуля) за $j + 1$ июня. Например, строка с номером `0` — прогноз продаж товара 0 на 1 июня, а строка `94` — прогноз продаж товара 3 на 5 июня. Всего в файле 600 строк с номерами от `0` до `599`. Все прогнозы — неотрицательные числа (нельзя продать отрицательное число товара). Товары упорядочены по возрастанию `ITEMID`. Пример такого файла — файл `baseline.csv`. Файл прогнозов нужно отправить в систему kaggle.

Правила kaggle.

- 3 попытки в день;
- Решения индивидуальные;
- Качество считается по метрике SMAPE (см. презентацию с семинара);
- До окончания соревнования доступны значения качества, посчитанные только на случайных 30% прогнозов. Значения отображаются в Public Leaderboard;
- После окончания соревнования становится доступным Private Leaderboard, в котором значения качества посчитанны на оставшихся 70% прогнозов;
- Баллы выставляются согласно Private Leaderboard;
- Для включения в Private Leaderboard можно выбрать две посылки;
- В Leaderboard должны отображаться ваши реальные имя и фамилия. В противном случае решение может быть не зачтено;
- Запрещается использование внешних наборов данных, не предоставленных на соревновании.

Методы решения.

В качестве решения можно использовать любые методы прогнозирования временных рядов, которые были рассмотрены на семинаре, а так же любые их модификации и комбинации. Кроме того, можно использовать любые модели, которые уже были рассказаны в каком-либо курсе из **обязательной программы**. Запрещается использовать какие-либо принципиально другие модели, например, нейронные сети. Запрет обусловлен тем, чтобы все участники были примерно в равном положении.

Базовое решение.

В ноутбуке `baseline.ipynb` содержится базовое решение, с помощью которого получается файл `baseline.csv`. Этот же файл соответствует строчке baseline в Leaderboard на kaggle. Задача — улучшить это решение или придумать другое решение, которое будет давать лучшее качество.

Формат решения.

Код решения должен быть на языке Python 3 или R в среде Jupyter Notebook (для R нужно установить IRkernel, чтобы работать с R в ноутбуках). Можно так же использовать комбинации двух языков, в частности, решение может быть последовательным применением нескольких ноутбуков. В решении можно использовать любые готовые реализации разрешенных методов. Каждый ноутбук должен быть сконвертирован в pdf. Кроме этого, нужно составить файл `N.txt`, в котором описать основные моменты решения понятным русским языком. Описание должно быть кратким по объему, но полным по содержанию.

На почту `probability.diht@yandex.ru` нужно выслать (`N` — номер из таблицы):

- `9.N.ipynb` и `9.N.pdf` — ноутбук и его pdf-версия (или `9.N.i.ipynb` и `9.N.i.pdf`, если ноутбуков несколько);

- `N.txt` — описание решения;
- `N.csv` — файл ответов.

Тема письма `"[СП17] Фамилия Имя - Задание 9"`. Квадратные скобки обязательны. Вместо `Фамилия Имя` нужно подставить свои фамилию и имя.

Все файлы должны соответствовать решению, на котором достигается наилучший результат, даже если оно не последнее. Если присланное решение не соответствует наилучшей попытке, ваш результат аннулируется, даже в случае если вы не можете восстановить все файлы наилучшей попытки. В реальных соревнованиях в таком случае вы можете лишиться большого денежного вознаграждения.

Совет: сохраняйте *все* файлы после каждой попытки, указывая номер попытки в названии файла или папки.

Баллы за задание.

- Решение лучше бейзлайна → 2 балла;
- Решение лучше бейзлайна не менее чем на 1% → 5 баллов;
- Решение лучше бейзлайна не менее чем на 5% и попадание в топ-10 Leaderboard → 10 баллов;
- Решение лучше бейзлайна не менее чем на 5% и попадание в топ-5 Leaderboard → 20 баллов;
- Решение лучше бейзлайна не менее чем на 5% и попадание в топ-3 Leaderboard → 20 баллов и +1 балл к оценке за семинары;
- Топ-1 Leaderboard → большая шоколадка.